(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 209 660 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
29.05.2002 Bulletin 2002/22

(51) Int Cl.[7]: **G10L 15/26**, G10L 13/04, H04M 3/493

(21) Application number: **01124578.4**

(22) Date of filing: **13.10.2001**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **23.11.2000 EP 00125606**

(71) Applicant: **International Business Machines Corporation**
Armonk, NY 10504 (US)

(72) Inventors:
- **Günther, Carsten, Dr.**
  69245 Bammental (DE)
- **Hänel, Walter**
  71088 Holzgerlingen (DE)
- **Schäck, Thomas**
  77855 Achern (DE)

(74) Representative: **Klein, Hans-Jörg**
IBM Deutschland GmbH
Intellectual Property Department
Pascalstrasse 100
70548 Stuttgart (DE)

(54) Voice navigation in web applications

(57) The present invention allows users to navigate in a web application or web pages using a combination of point-and-click and voice-input. At each point of the dialog, the user can use the standard point-and-click interface to perform context-dependent actions or use speech input to navigate and operate in the global application context alternatively. The voice input uses a voice navigation component which builds an interface to the installed recognition and synthesis engines. The point-and-click and the voice navigation component will be loaded automatically with the initial web page of a web application. Grammars for recognizing vocabulary related to that web application will be provided with the voice navigation component. The present invention combines the advantages of a context- dependent point-and-click User Interface with those of a context-independent speech-input interface. It is an approach to enhance we browsers towards multi-modal interfaces.

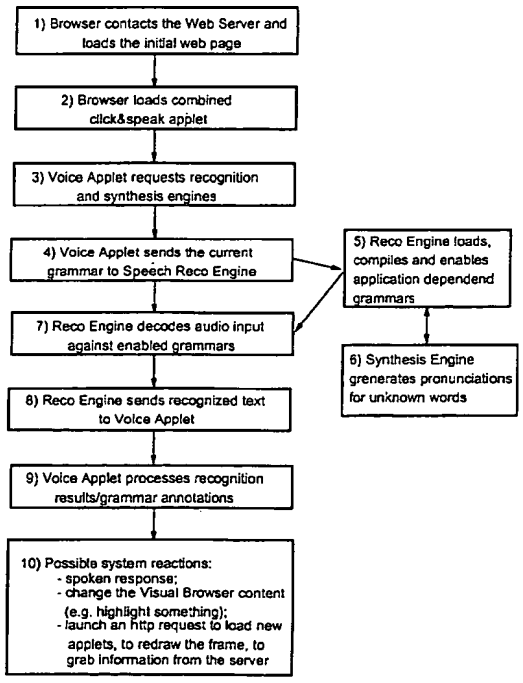

FIG. 4

EP 1 209 660 A2

Description

[0001] The present invention discloses a system and method for gathering information by voice input, especially a system and method for context-independent navigation in web applications or related web pages using voice input.

Field of the Invention

[0002] The present invention relates generally to a voice-driven system and method for gathering information accessible via a network, especially Intranet or Internet.

Description of the Related Art

[0003] Hypertext systems are rapidly gaining increasing significance in many areas of data and communications technology. The important examples that have already been realized are:

[0004] Typical hypertext help systems and hypertext documentation for software applications (for example, under graphics operating systems for personal computers), wherein the user can usually navigate within single hypertext documents that are stored as data files on a single computer, as well as the World Wide Web (WWW), a worldwide hypertext network based on the Internet that makes it possible for the user to navigate through a plurality of hypertext documents linked to one another that cite one another (i.e., reference one another) and that are generally stored on a great number of computers in the network at a great distance from one another. Hypertext documents thereby generally contain information in the form of text, digital images or audio or video data or combinations thereof.

[0005] A significant, characteristic feature of all hypertext systems is the possibility of navigation. In addition to containing the actual text of the document, a hypertext document contains special character sequences that can also be a component part of the actual text and that are usually referred to as links or hyper-links and that serve the purpose of hypertext navigation. Normally, these character sequences are specially marked, for example by being displayed in a different color or are emphasized in some other way, to distinguish the special character sequences from the ordinary text of the document. When a user of a hypertext system selects such a link, usually with a short click with the mouse or other pointing device, the hypertext system reacts to this instruction in that it displays the part of the same hypertext document associate with this character sequence (or link) or the system displays a different hypertext document. Other possible reactions to the selection of the link are opening up a connection to a different computer, for example to an on- line data bank, starting another application program, opening another data file, initiating a data processing process or a combination of such possible reactions. In addition thereto, hypertext systems usually also execute other instructions that are not associated with the character sequences (links) in the hypertext documents such as, for example, paging through documents that are already displayed or through document parts (for example, pages of the document), storing hypertext pages in what are referred to as hot lists, retrieving or paging through pages stored in hot lists, reloading images, etc. These instructions are normally input in the way typical for graphic user interfaces, for example with the mouse or other pointed device. There are a number of possible applications of hypertext-based systems wherein the traditional way of inputting instructions or of activating links is considered to be disturbing, undesirable or even impossible. This, for example, is the case when the user is impeded, his hands are busy with managing other jobs or when the ambient conditions forbid the employment of traditional input devices. Voice recognition is available here as a simple, natural type of input that assumes less expertise on the part of the user than other input means. The integration of traditional, acoustic voice recognition systems, i.e. systems for recognizing spoken language, with hypertext systems, which are also known as "viewer" or "browser" systems, are opposed by technological difficulties. The voice recognition system, namely, must be in the position to recognize every word that can occur as a link in a hypertext document. Because practically every word in the text can also be a hyper-link, extremely large dictionaries would be required for this purpose, and these large dictionaries would reduce the processing speed and the recognition performance of this system to an unjustifiable extent. Even if the employment of extremely large dictionaries were possible, the numerous coining of new words and proper names could not be recognized, these new words being so typical of many hypertext applications, specifically, however, for hypertext networks such as the World Wide Web.

[0006] US 6020135 discloses a hypertext navigation system for voice controlled navigation wherein a dictionary is provided which includes probability models for spoken words. The dictionary and probability model, which includes phoneme sequences to be matched to the spoken words, is generated in the user's system during access to the hypertext document in the run time version. An off- line version provides a dictionary and probability model that is generated by the author, for example, of the hypertext document, is stored on the server and is forwarded to the user system when the document is accessed by the user. The dictionary and probability model correspond to the hypertext elements that are in the hypertext document which is being accessed by the user. Accordingly, the dictionary and probability model are discarded and the next dictionary and probability model obtained as the next hypertext document is accessed. Storage of recent or important dictionaries and probability models are also provided.

[0007] A disadvantage of that system is that the voice

recognition is mainly restricted to the hyperlinks used in the hyertext document which is being accessed by the user. Other hyperlinks which are not visible on the hypertext document being accessed cannot be recognized. Furthermore, fill out forms cannot be handled by that prior art system.

[0008] It is therefore object of the present invention to provide a hypertext navigation system combining the advantages of point and click hypertext navigation system with prior art voice controlled hypertext navigation system by avoiding their disadvantages.

[0009] This object is solved by the features of the independent claims. Further preferred embodiments are laid down in dependent claims.

[0010] The present invention allows users to navigate in a web application or web pages using a combination of point-and-click interaction and voice-input and voice-output interaction. At each point of the dialog, the user can use the standard point-and-click interface to perform context-dependent actions or use speech input to navigate and operate in the global application context alternatively. The voice input uses a voice navigation component which builds an interface to the installed recognition and synthesis engines. The point-and-click and the voice navigation component will be loaded automatically with the initial web page of a web application. Grammars or language models for recognizing vocabulary related to that web application will be provided with the voice navigation component. The present invention combines the advantages of a context-dependent point-and-click user interface with those of a context-independent speech-input interface.It is an approach to enhance web browsers towards multi-modal interfaces.

[0011] The present invention will be described in more detail using a preferred embodiment with Figures, where

FIG.1 illustrates the architecture in which the present invention may be used preferably

FIG.2 illustrates a preferred embodiment of the present invention used in the architecture according to FIG.1

FIG.3 illustrates the inventive method according to FIG.2

FIG.4 illustrates a flow chart with the inventive steps for carrying out the present invention

FIG 5 illustrates a preferred user interface for activating the inventive point-and-click and voice navigation component (applet) by the user.

FIG.6 illustrates the relationship of the interfaces between the inventive voice navigation component (applet) and the voice recognition and speech synthesis component

FIG.7 illustrates the use of the vocabularies by the inventive voice navigation component (applet)

FIG.8 illustrates the logical structure of a speech recognition system with its major components and how the inventive voice navigation component (applet) interacts with the speech recognition system

[0012] In FIG.1 the basic architecture is shown in which the present invention may be implemented preferably. The basic architecture may be a client-server architecture. On the client side following standard components are at least installed:

audio output device (e.g. loud speaker or head phones)(2) microphone (4) web browser (e.g. Netscape (6))

speech recognition and speech synthesis system (e.g. IBM Via Voice (8) and IBM Via Voice Outloud (10).

[0013] The heart of the speech recognition system is known as speech recognition engine. The speech engine recognizes speech input and translates it into text that an application understands. The application decides what to do with the recognized text. Speech-aware applications (18) access the speech engine and various speech resources through a speech recognition API (Application Programming Interface).

[0014] The speech engine may use following resources to process spoken words:

User's language of origin
Grammars

[0015] The language of origin is the language used by the speaker.

[0016] Each language can include several different grammars. A grammar is a set of vocabularies, pronunciations, and word usage models designed to support the application. The grammar is used by the speech engine to decode speech for the application. The application specifies the set of active words by activating one or more grammars.

[0017] On the server side following standard components are preferably installed:

Web Server or HTTP-Server (14)
one or more Web applications or servlets (18)
an application server or/and a data base (16)

[0018] FIG.2 illustrates the implementation of the present invention into a client-server architecture as shown in FIG.1. The speech recognition and synthesis system are available to signed Java applets.

[0019] The main component of the present invention

is the voice navigation component (applet). The voice navigation component (applet) (2) performs the following major steps:

- locate, select, and initialize a speech recognition engine and speech synthesis engine
- define, enable, and disable decoding grammars
- processing of the recognition results (e.g. launch HTTP request, initiate spoken words, play back prerecorded prompts).

[0020] It is possible to use general grammars or language models that are available at the client side (60).

[0021] Usually they are installed along with the general speech recognition engine (10). Furthermore it is required to upload application-dependent or so called information-dependent grammars from the server to the client (60). These grammars specify the recognition vocabulary for navigating within related web pages or web pages belonging to a web application or related web applications. The point-and click navigation component (applet 4) presents visible and activable menu items or fillable fields. This method is user unfriendly or highly structured user interfaces to web applications (servlets-80) because it requires a lot of clicks to step down through a menu structure or to switch into a new menu context. Therefore, it is much more user friendly to use the more general inventive voice navigation component (applet) (2). Possible input values (spoken words) to select links, menu items or to fill out forms in a visible web page or non-visible web pages can be defined via grammars. It is therefore not necessary to restrict valid input values to visible links. Additionally, it is also possible to speech-enable out of context or more general links as shortcuts to avoid time consuming menu navigation.

[0022] A further component of the present is the conventional point-and- click navigation component (applet 4) as used in existing prior art systems (mouse systems). The point-and-click component (applet PACNA) allows to load new web pages by pointing and clicking hyperlinks displayed in HTML documents.

[0023] Both components (2; 4) are originally stored on the server system
and preferably loading of a initial web page (6) from the server (40) to the client will initiate a loading of both components automatically. As far the application dependent grammars are laid down in separate applets or files on the server (40) they may be loaded in conjunction with the initial web page (6) containing links (reference information/ URIs) to the respective application grammar. Another implementation may be that the grammars are part of the voice navigation component (applet).

[0024] The point-and-click navigation component (applet 4) and the voice navigation component (applet) (2) process the respective user input to produce an HTTP-request required to load a new web page.

[0025] The user may select between both components(2,4) alternatively by clicking the appropriate applet symbol displayed in the GUI on the client display provided by the web-application (servlet) preferably.

[0026] Further standard components on the server side may be a Web Server (e.g. IBM HTTP-Server; 70), an Application Server (e.g. IBM Websphere; 65) and a database (90). Web Server and Web browser communicates with each other and servlets(80) and applets (2,4) are stored on the server (40). The servlets will be executed on the server side and the applets will be executed on the client side.

[0027] On the client side a Java Virtual Machine (100) must be available for processing the Java-applets.

[0028] FIG.3 illustrates the basic structure of the voice navigation component (applet) according to FIG 2.

[0029] The voice navigation component (applet 2) which has been loaded from the server (40) to the client (60) uses the Client's voice recognition system(10) via the JVM (100). It connects to the installed recognition and synthesis systems, grammars or language models for the web-applications to be accessed (servlets; 80) are enabled or loaded and prompts are played. The voice navigation component (applet 2) passes audio input to the speech recognition engine(10) to decode against enabled grammars .The recognition result contains recognized words/phrases and grammar annotations. The voice navigation component (applet 2) specifies the processing of the recognition result. Relevant information of the result is extracted and is sent to the server (40), e.g. to a servlet. The server (40) may further process the request and as result returns a response with a new web page (6) for example. Possible reactions may be change the browser content, launch a http request to load new Web page, grab information from the server and to initiate a server-based transaction. The processing of the recognition result may be done either in the client (60) or in the server (40) or the processing may be distributed partly to the client (60) and the server (40). For example the semantic processing of the speech input may be distributed between client (60) and server (40). A possible implementation may be that the initial signal processing may be accomplished by a signal processing applet on the client side, the feature vector is sent via the network to the server side, and the speech recognition is made on the server side.

[0030] FIG.4 describes in the form of a flow chart the inventive process steps of voice-activated navigation according to the present invention.

1. Browser contacts the Web Server and loads an initial web page (2).

2. Browser loads combined point-and-click and voice navigation component (applet). The initial web page contains reference information/links (URIs-) to the point-and-click and voice navigation component (applet). The Browser evaluates the URIs and loads the respective component(applet)s (4).

3. Voice navigation component (applet) requests recognition and synthesis engines. The Java Virtual Machine processes both component(applet)s. The voice navigation component (applet) initializes the voice driven user interface. It locates, selects, and creates a speech recognition engine and a speech synthesis engine. The speech recognition engine is responsible for processing audio input to the Browser whereas the speech synthesis engine creates spoken words (6).

4. Voice component(applet) sends the current vocabularies to the speech recognition engine (8).

The recognition of incoming speech is grammar driven. The actually valid grammar is defined in applets which will be loaded with voice navigation component (applet). The grammar will contain words/phrases matching words/phrases visible in the browser window. Furthermore, the voice navigation component (applet) may activate additional words/phrases that do not match expressions in the browser window. The present invention allows to enable words/phrases from a broader context, namely to enable word phrases for navigating within related web pages or web pages belonging to a web application or related web applications, e.g. general navigation commands, help commands, additional sub menu items and so on (information-dependent grammars). This allows direct voice driven jumps into application's sub menu and overcomes the cumbersome approach of clicking through endless menu lists and check boxes

5. Speech recognition engine loads, compiles and enables information/application-dependent grammars (10).
The recognition engine enables the defined grammars. It is possible to enable multiple grammars for recognizing a broad scope speech. Within the grammars the actual valid recognition vocabulary is defined (10).

6. Synthesis engine generates pronunciations for unknown words (12). A speech recognition engine comes along with a basic vocabulary and attached pronunciations. But an application can contain unknown words. The recognition engine sends a request to the synthesis engine to generate missing pronunciations. These words are then added to the actual enabled words.

7. Speech recognition engine decodes audio input against enabled grammars(14). Incoming audio input is routed to the speech recognition engine. The speech recognition engine decodes against the enabled grammars.

8. Speech recognition engine sends recognized text to voice navigation component (applet) (16).

The recognition result contains recognized words/phrases and grammar annotations. Grammar annotations represent return values of recognized grammar phrases and allow a flexible processing of recognition results. Misrecognitions (e.g. incomplete phrases, low audio input level) have to handled by the voice navigation component (applet).

9. Voice navigation component (applet) specifies the processing of the recognition result (18).

10. Possible reactions are:

- spoken response
- change of the browser content
- launch a http request to load a new application/ applet or web page, to redraw the content frame, to grab information from a server, to initiate a server-based transaction (20).

[0031] FIG.5 illustrates an example of user interface for the point-and- click navigation component (applet) and the voice navigation component (applet) preferably used in the present invention.

[0032] The part of user interface of the voice navigation component (applet) represents several options (6) for enabling or activating different grammars. For example option 0-3 allows to activate grammars which are restricted to recognize visible links only and option 2-2 allows to activate grammars, information dependent grammars, which opens the possibility to speech-enable out of context or more general links by avoiding time consuming navigation procedures.

[0033] FIG.6 illustrates the advantages of the present invention with respect to a stock brokerage application for buying stocks of a certain company via Internet. Starting with the home page of the application the user has to click down from the link "customer function" to the data entry field indicated by the arrow. Then he has to input the appropriate data in the data field by typing in information. By using the present invention the user can voice-driven navigate directly from the link "customer function" to the desired data entry field and can also fillout the data entry field by voice without typing in any information.

[0034] This is realized by a grammar (applet) recognizing general navigation commands, help commands, additional sub menu items and so on contained in that brokerage application.

[0035] FIG.7 illustrates the relationship of the interfaces by a specific implementation of the present invention into IBM's Via Voice Speech Recognition Engine (8) and Text -to-Speech Engine (10).

[0036] The application programming interface to the IBM Via Voice Engine is SMAPI (12). It supports:

- verifying the API version
- Establishing a database session query system parameter
- Establishing a recognition session
- Setting up vocabularies
- Setting speech engine parameters
- Processing speech input
- Adding new words to the user vocabulary
- Handling errors
- Disconnecting from the speech engine
- Closing a speech session

[0037] The SMAPI (8) is provided as a DLL which may be linked into the voice navigation component (applet) (14).

[0038] The application programming interface to the IBM Via Voice Text-to-Speech Engine (10) is called SAPI (16). The Text-to-Speech Engine uses the following resources to translate text into synthesized speech:

- user dictionaries
- special words
- abbreviations
- roots.

[0039] The SAPI is provided as a DLL which may be linked into the voice navigation component (applet) (14).

[0040] As far as the voice navigation component (applet) is written in the program language Java an additional Java API is layered between SMAPI and SAPI (12,16) and the voice navigation component (applet) (14). The Java API may also be provided as a DLL which may be linked into the voice navigation component (applet- not shown)

More detailed information about the IBM ViaVoice programming interfaces are accessible via http://w3.speech.ibm.com/tkdoc/ViaVoice/proguide/pgmguiO3.ht m

Claims

1. A client system for gathering information via a network by voice input, comprising:

   a speech recognition engine (10) installed on said client system (60);

   communication component(12) installed on said client system for establishing communication to a communication component on a server system (70) providing access to information (6) stored on said server

   additionally **characterized by**

   a voice navigation component(2) providing information-dependent grammars from said server (40) to said speech recognition engine (10) via said communication component(12) based on initial information loaded from said server to said client and processing the results of said speech recognition system (10).

2. System according to claim 1, wherein said speech recognition engine (10) further includes speech synthesis engine.

3. System according to claim 1, wherein said communication component (12) and said voice navigation component forms an integral component.

4. System according to claim 1, wherein said communication component is a browser.

5. System according to claim 1, wherein said voice navigation component (2) has the following functionality:

   locate, select, and initialize a speech recognition engine and speech synthesis engine

   enable, and disable information-dependent grammars

   processing of the recognition results

6. System according to claim 1, wherein said network is an Intranet or an Internet.

7. Client-Server System comprising
   a client (60) having
   a speech recognition engine and speech synthesis engine (10)
   client communication component(12) for establishing communication with a server (40)
   a voice navigation component(2) providing information-dependent grammars from said server to said speech recognition engine via said client communication component based on initial information loaded from said server to said client and processing the results of said speech reognition engine.
   a server (40) having
   server communication component (70) for establishing communication with a client (60)
   a voice navigation component (2) providing information-dependent grammars from said server to said speech recognition engine based on initial information loaded from said server to said client and processing the results of said speech reognition engine, wherein said voice navigation component is available for download to and execution on said client (60)
   information-dependent grammars available for download to and execution on said client.

8. Method for gathering information via system according to claim 1 comprising the following steps:

   loading an initial information from said server to said client by means of said communication components

   automatically loading an information-dependent grammar to said client by using access information contained in said initial information and automatically providing said information-dependent grammar to said speech recognition engine for recognizing spoken words defined by said information-dependent grammar

   sending result of said speech recognition engine to said Voice Navigation component

   processing result of said speech engine said Voice Navigation component.

9. Method according to claim 8, wherein said information-dependent grammar defines possible input values of web related web pages or web pages belonging to a web application or related web applications.

10. Method according to claim 8, wherein said initial information may be any web page made available by said server.

11. Method according to claim 8, wherein said initial web page contains a reference to said voice navigation component stored on said server.

12. Method according to claim 8, wherein each initial web page contains a reference to a point-and-click component stored on said server.

13. Method according to claim 8, wherein loading of said initial web page is accompanied by the following steps:

   automatically identifying reference information in said initial web page for accessing said voice navigation and point-and-click component and automatically loading said voice navigation and point-and-click component from said server to said client using said reference information.

14. Method according to claim 8, wherein loading of said initial web page is accompanied by the further steps:

   automatically identifying reference information to information - dependent grammars in said initiating web page

   automatically loading said identified information-dependent grammar to said client

   providing access of said speech recognition engine to said information-dependent grammar by said voice navigation component.

15. Method according to claim 8, wherein said voice navigation and said point-and-click component having an common user-interface with options selectable by said user.

16. Method according to claim 15, wherein said user interface for the voice navigation component offers options for selecting information-dependent grammars stored on said server.

17. Method according to claim 8, wherein said voice navigation component processes following results from said speech recognition and synthesis engine:

   - spoken response
   - change of the browser content
   - launch a http-request to load a new application/ applet or web pages,
   - redraw the content frame,
   - to grab information from a server,
   - to initiate a server-based transaction

18. Computer program product stored on a computer-readable media containing software for performing of the method according to 8 to 17 when said program is executed on a computer.

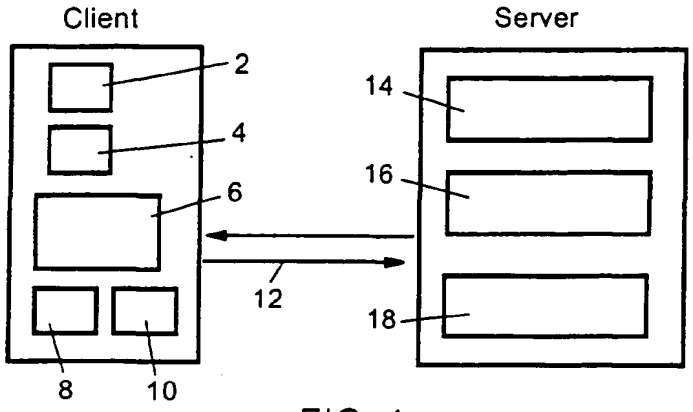
Client
Server
2
4
6
8
10
12
14
16
18

FIG. 1

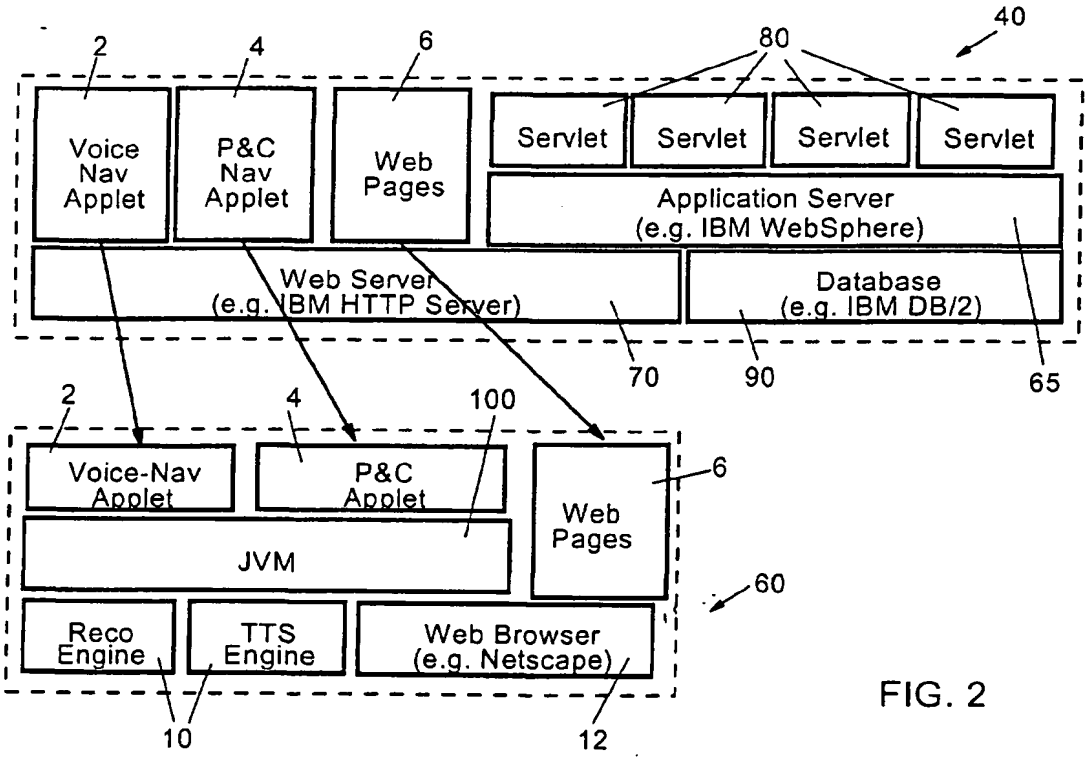
40
2
4
6
80
Voice Nav Applet
P&C Nav Applet
Web Pages
Servlet
Servlet
Servlet
Servlet
Application Server (e.g. IBM WebSphere)
Web Server (e.g. IBM HTTP Server)
Database (e.g. IBM DB/2)
70
90
65
2
4
100
Voice-Nav Applet
P&C Applet
Web Pages
6
JVM
60
Reco Engine
TTS Engine
Web Browser (e.g. Netscape)
10
12

FIG. 2

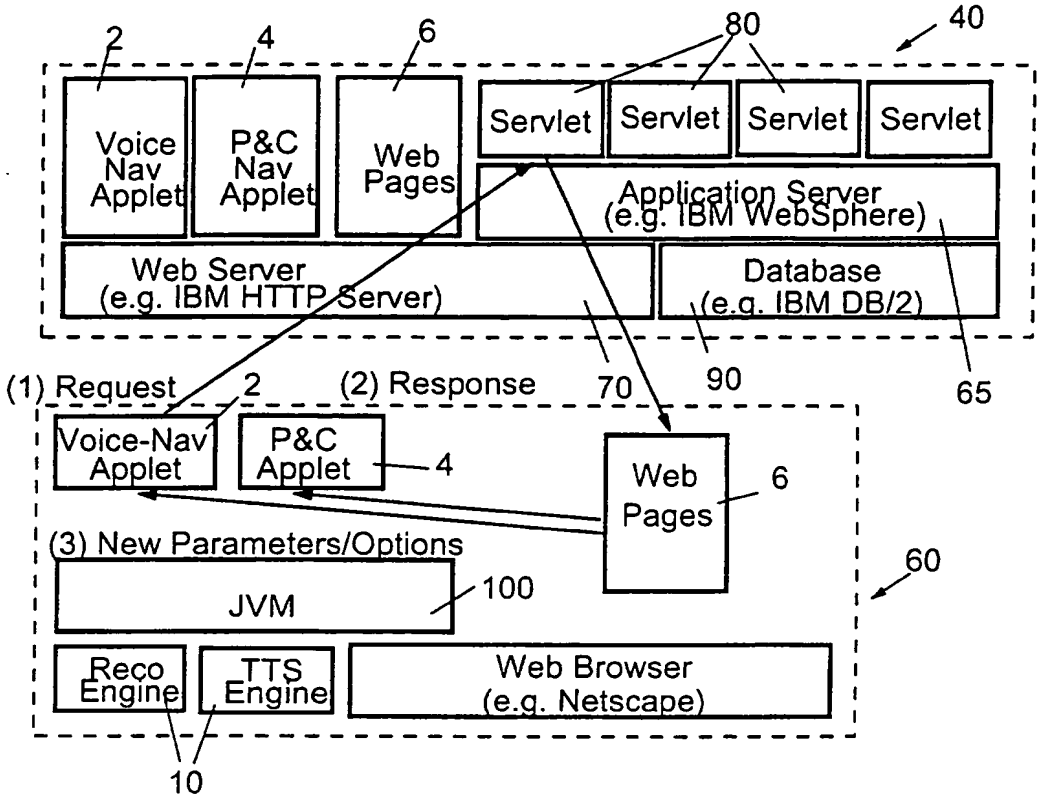
2
4
6
80
40
Voice Nav Applet
P&C Nav Applet
Web Pages
Servlet
Servlet
Servlet
Servlet
Application Server (e.g. IBM WebSphere)
Web Server (e.g. IBM HTTP Server)
Database (e.g. IBM DB/2)
(1) Request
2
(2) Response
70
90
65
Voice-Nav Applet
P&C Applet
4
Web Pages
6
(3) New Parameters/Options
60
JVM
100
Reco Engine
TTS Engine
Web Browser (e.g. Netscape)
10

FIG. 3

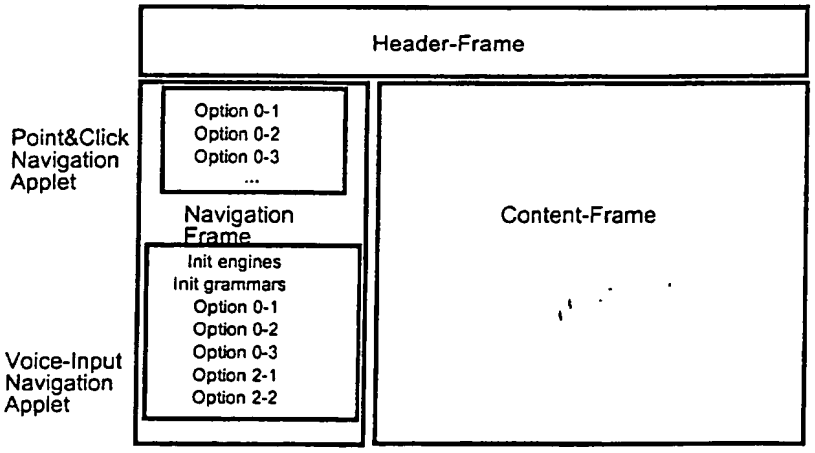
Header-Frame
Point&Click Navigation Applet
Option 0-1
Option 0-2
Option 0-3
...
Navigation Frame
Content-Frame
Init engines
Init grammars
Option 0-1
Option 0-2
Option 0-3
Option 2-1
Option 2-2
Voice-Input Navigation Applet

FIG. 5

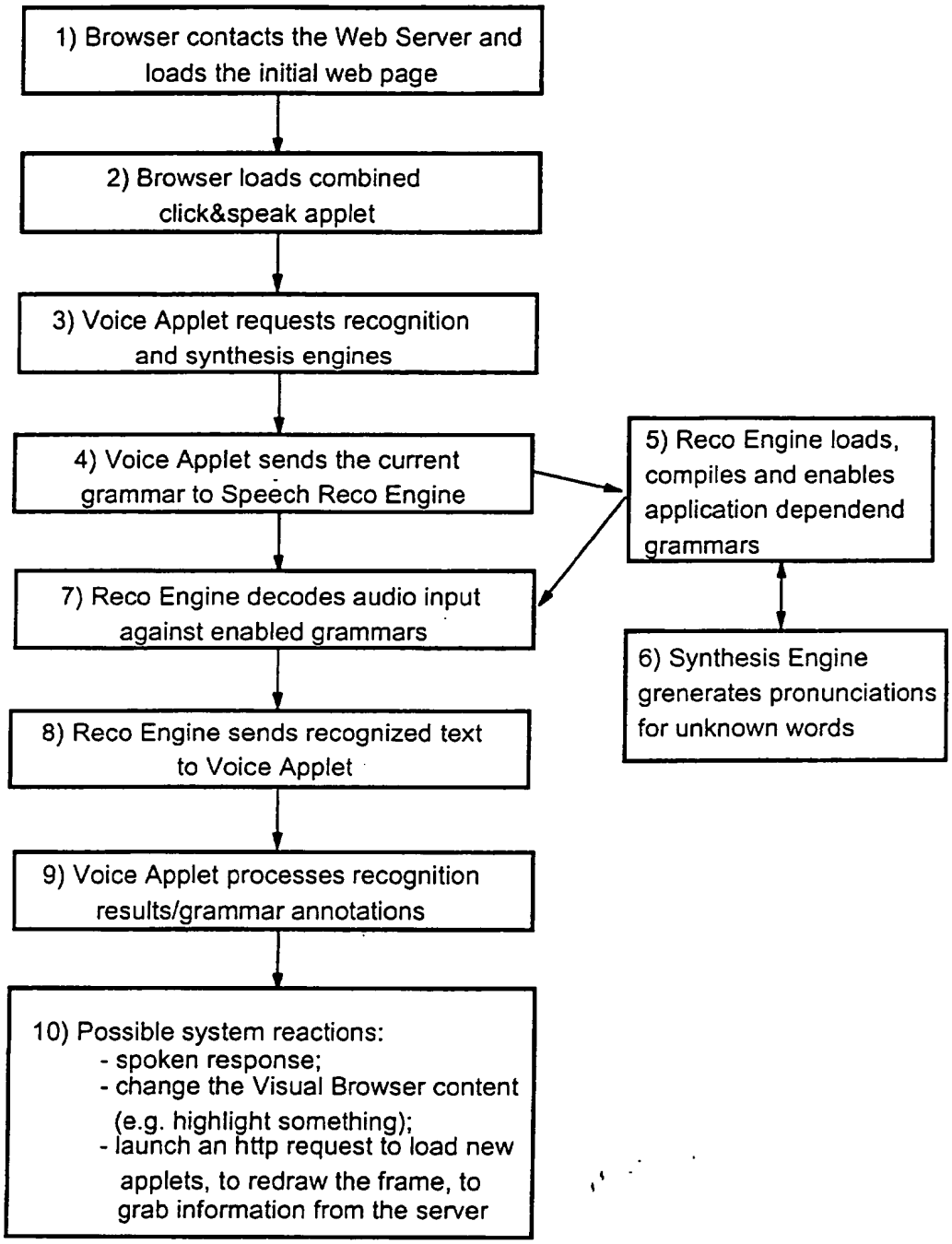
1) Browser contacts the Web Server and loads the initial web page
2) Browser loads combined click&speak applet
3) Voice Applet requests recognition and synthesis engines
4) Voice Applet sends the current grammar to Speech Reco Engine
5) Reco Engine loads, compiles and enables application dependend grammars
7) Reco Engine decodes audio input against enabled grammars
6) Synthesis Engine grenerates pronunciations for unknown words
8) Reco Engine sends recognized text to Voice Applet
9) Voice Applet processes recognition results/grammar annotations
10) Possible system reactions:
- spoken response;
- change the Visual Browser content (e.g. highlight something);
- launch an http request to load new applets, to redraw the frame, to grab information from the server

FIG. 4

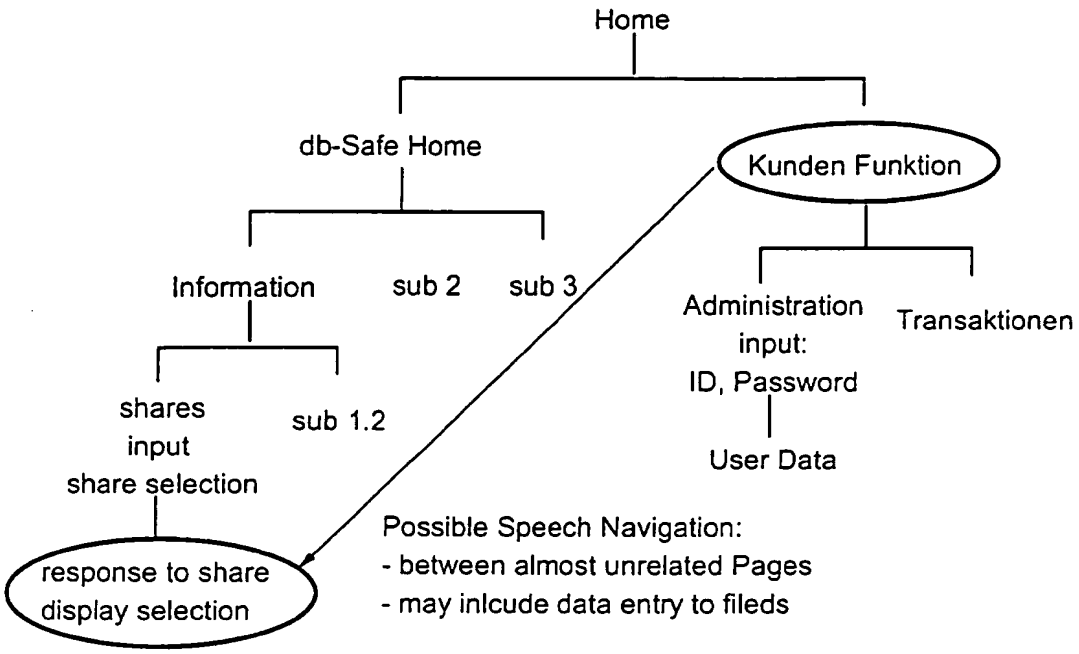
Home
db-Safe Home
Kunden Funktion
Information
sub 2
sub 3
Administration
input:
ID, Password
Transaktionen
shares
input
share selection
sub 1.2
User Data
response to share
display selection
Possible Speech Navigation:
- between almost unrelated Pages
- may inlcude data entry to fileds

FIG. 6

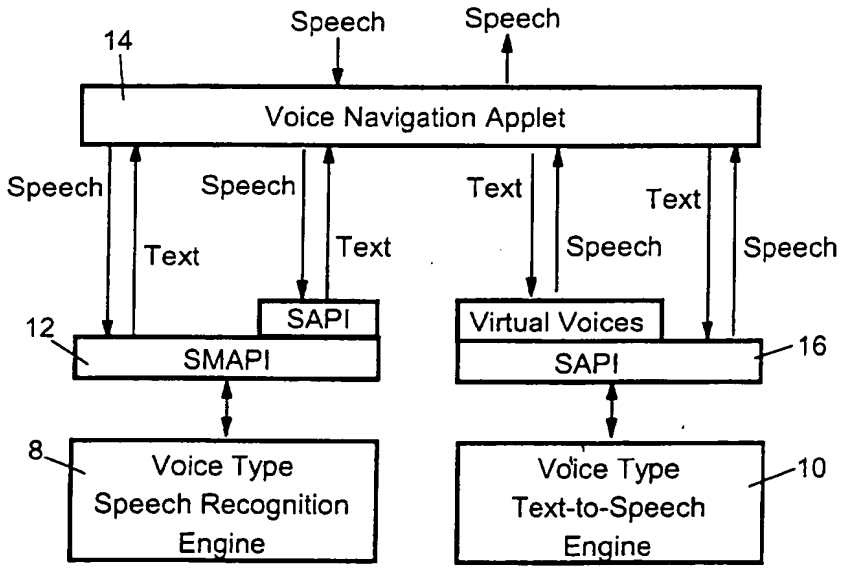
14
Speech
Speech
Voice Navigation Applet
Speech
Text
Speech
Text
Text
Speech
Text
Speech
12
SAPI
SMAPI
Virtual Voices
SAPI
16
8
Voice Type
Speech Recognition
Engine
Voice Type
Text-to-Speech
Engine
10

FIG. 7